대한민국특허청
KOREAN INTELLECTUAL PROPERTY OFFICE

별첨 사본은 아래 출원의 원본과 동일함을 증명함.

This is to certify that the following application annexed hereto is a true copy from the records of the Korean Intellectual Property Office.

출 원 번 호 : 10-2003-0072244
Application Number

출 원 년 월 일 : 2003년 10월 16일
Date of Application OCT 16, 2003

출 원 인 : 한국전자통신연구원
Applicant(s) Electronics and Telecommunications Research Insi

2003 년 11 월 24 일

특 허 청
COMMISSIONER

【서지사항】

【서류명】	특허출원서
【권리구분】	특허
【수신처】	특허청장
【제출일자】	2003.10.16
【발명의 명칭】	백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법
【발명의 영문명칭】	Semi-Automatic Construction Method for Knowledge of Encyclopedia Question Answering System
【출원인】	
【명칭】	한국전자통신연구원
【출원인코드】	3-1998-007763-8
【대리인】	
【성명】	권태복
【대리인코드】	9-2001-000347-1
【포괄위임등록번호】	2001-057650-1
【대리인】	
【성명】	이화익
【대리인코드】	9-1998-000417-9
【포괄위임등록번호】	1999-021997-1
【발명자】	
【성명의 국문표기】	왕지현
【성명의 영문표기】	WANG,Ji Hyun
【주민등록번호】	740116-1074237
【우편번호】	305-345
【주소】	대전광역시 유성구 신성동 119-11 선경빌라 301호
【국적】	KR
【발명자】	
【성명의 국문표기】	정의석
【성명의 영문표기】	CHUNG,Eui Sok
【주민등록번호】	730203-1347617
【우편번호】	305-345
【주소】	대전광역시 유성구 신성동 208-3 은지빌라 101호
【국적】	KR

【발명자】

【성명의 국문표기】	장명길
【성명의 영문표기】	JANG,Myung Gil
【주민등록번호】	650522-1095018
【우편번호】	305-335
【주소】	대전광역시 유성구 궁동 다솔아파트 103동 504호
【국적】	KR
【심사청구】	청구
【취지】	특허법 제42조의 규정에 의한 출원, 특허법 제60조의 규정에 의한 출원심사 를 청구합니다. 대리인 권태복 (인) 대리인 이화익 (인)

【수수료】

【기본출원료】	20	면	29,000 원
【가산출원료】	9	면	9,000 원
【우선권주장료】	0	건	0 원
【심사청구료】	14	항	557,000 원
【합계】	595,000 원		
【감면사유】	정부출연연구기관		
【감면후 수수료】	297,500 원		

【기술이전】

【기술양도】	희망
【실시권 허여】	희망
【기술지도】	희망
【첨부서류】	1. 요약서·명세서(도면)_1통

【요약서】

【요약】

본 발명은 지식베이스의 구조를 설계함에 있어 백과사전의 내용을 기반으로 개념 중심의 체계적인 템플릿을 설계하고, 백과사전의 개요 정보 및 본문으로부터 표제어와 관련된 중요한 사실 정보를 자동으로 추출하여 질의응답시스템의 지식베이스를 반자동으로 구축하는 방법에 관한 것이다.

본 발명은, 각 표제어에 대해 다수의 템플릿들과 관련 속성들로 지식베이스 구조를 설계하는 단계와, 백과사전의 개요정보로부터 표제어와 그 속성이름 및 속성값들을 추출하는 단계와, 문장분석을 통해 얻어지는 어절단위 토큰열의 의존관계를 기반으로 백과사전의 본문으로부터 그 표제어에 대한 속성이름 및 속성값들을 추출하는 단계와, 각 표제어별로 상기 추출된 구조정보 및 비 구조정보를 지식베이스의 해당 템플릿 및 해당 속성에 저장하여 지식베이스를 구축하는 단계로 이루어진다.

【대표도】

도 2

【색인어】

질의응답시스템, 지식베이스, 템플릿, 의존관계, 백과사전, 최대 엔트로피 모델

【명세서】

【발명의 명칭】

백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법{Semi-Automatic Construction Method for Knowledge of Encyclopedia Question Answering System}

【도면의 간단한 설명】

도 1은 본 발명의 정보추출 대상인 백과사전의 예시 도면.

도 2는 본 발명에 따른 백과사전 지식베이스 구축 시스템의 개념도.

도 3은 본 발명에 따른 지식베이스 템플릿에 대한 예시 도면.

도 4는 본 발명에 따른 템플릿의 세분류에 대한 예시 도면.

도 5는 본 발명에 따른 구조 정보 추출과정을 도시한 흐름도.

도 6은 본 발명에 따른 비구조 정보 추출과정을 도시한 흐름도.

도 7은 각 격 조사에 따른 격 정보를 보여주는 도표도.

도 8은 본 발명에 따른 비구조 정보 추출예를 보여주는 도면.

도 9는 본 발명에 따른 속성태그 의미에 대한 도표도.

<도면의 주요부분에 대한 부호의 설명>

100: 표제어 101: 개요 정보

105: 본문 201: 구조정보 추출 모듈

202: 비 구조정보 추출 모듈

203: 지식베이스 600: 통계모델 DB

【발명의 상세한 설명】

【발명의 목적】

【발명이 속하는 기술분야 및 그 분야의 종래기술】

<15> 본 발명은 질의응답 시스템의 지식베이스 구축에 관한 것이며, 보다 상세히는 지식베이스의 구조를 설계함에 있어 백과사전의 내용을 기반으로 개념 중심의 체계적인 템플릿을 설계하고, 백과사전의 개요 정보 및 본문으로부터 표제어와 관련된 중요한 사실 정보를 자동으로 추출하여 지식베이스에 저장하며, 특히 비구조 정보 추출시 문장내 의존관계 분석과 최대 엔트로피 모델을 이용하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법에 관한 것이다.

<16> 인터넷 상의 정보검색 시스템은 일반적으로 키워드 단위의 불리언 매칭에 의한 검색이 많이 사용되고 있으며 백과사전의 검색 서비스에서도 널리 적용되고 있다.

<17> 하지만, 종래 일반적인 백과사전 검색 서비스들은 사용자로부터 표제어를 입력받아 해당 표제어의 본문 내용을 브라우징하는 정도에 그치고 있다. 차세대 정보검색 서비스로서 질의응답 서비스를 적용하는 사례가 간간히 나타나고 있지만 사용자의 질의 요구에 만족할 만한 풍부한 답변을 제공하지 못하는 실정이다.

<18> 이것은 웹 문서나 백과사전 문서의 내용이 매우 방대할 뿐만 아니라 다양하고 복잡한 자연어 텍스트로 작성되어 있어서, 이들로부터 가치있는 유효 정보를 추출하여 색인하는 것이 쉽지 않기 때문이다.

<19> 또한, 종래의 백과사전 질의응답 시스템들은 사용자의 자연어 질의에 대한 정답을 제시하기 위해 지식베이스를 구축함에 있어, 자동화된 방법을 사용하지 않고 일반적으로 사람에 의

해 수동적으로 질문과 정답을 구축하기 때문에 지식베이스 구축에 많은 노력과 비용이 소모된다. 그리고, 이와 같은 지식베이스의 구조는 질문에 대한 답을 미리 예상하여 저장해 놓는 형태이기 때문에 지식베이스의 구조가 단순하며 체계적이지 못할 뿐만 아니라 유연성 및 활용면에서도 떨어지는 단점이 있다.

<20> 한편, 최근 반자동 지식베이스 구축과 관련하여 정보추출 분야에 대한 연구가 각국에서 진행되고 있다. 영어권 언어의 경우 한국어와 달리 문장의 형식이 보다 정형화되어 있어서 패턴의 적용이 수월한 편이기 때문에 Autoslog, Whisk, Crystal 등이 영어권 언어에 대해 비교적 양호한 성능을 보이고 있지만, 한국어 등과 같이 동일한 의미의 자연어 문장에서도 매우 다양한 형태로 표현이 가능할 경우는 이러한 방식을 적용하는 것은 적절하지 않으며 결국 지식베이스에 많은 패턴들을 일일이 구축해야 하는 단점이 있다.

【발명이 이루고자 하는 기술적 과제】

<21> 따라서, 본 발명은 상술한 종래의 문제점을 해결하기 위한 것으로서, 본 발명의 목적은 백과사전 질의응답 시스템에 있어서 지식베이스의 구조를 보다 체계적으로 설계하는 방법을 제공하며, 백과사전의 개요 정보 및 본문으로부터 표제어와 관련된 중요한 사실 정보를 자동으로 추출하여 지식베이스에 저장함으로써 지식베이스 구축에 따른 시간 및 비용의 부담을 줄일 수 있고, 질의응답 시스템의 효율 및 성능을 제고시키며 지식베이스의 완성도를 향상시키는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법을 제공하는데 있다.

<22> 또한, 본 발명은 백과사전의 본문에서 비구조 정보를 추출시 통계 모델을 적용함에 있어서 기존 접근방법과 같이 단순한 좌우 주변 문맥을 학습 자질로 하는 것이 아니라 문장내의 의존관계를 분석하여 학습 자질을 추출하기 때문에 비교적 자유로운 어순을 따르는 한국어를 포

함한 비영어권 언어에 대해서도 용이하게 적용될 수 있는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법을 제공하는데 있다.

<23> 상기 본 발명의 목적을 달성하기 위한 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법은, 각 표제어에 대해 다수의 템플릿들과 각 템플릿에 대한 다수의 관련 속성들로 지식베이스의 구조를 설계하는 지식베이스 설계단계; 백과사전의 개요정보로부터 표제어와 그 속성이름 및 속성값들을 추출하는 구조정보 추출 단계; 백과사전의 본문으로부터 그 표제어에 대한 속성이름 및 속성값들을 추출하는 비 구조정보 추출 단계; 및 각 표제어별로 상기 추출된 구조정보 및 비 구조정보를 지식베이스의 해당 템플릿 및 해당 속성에 저장하는 지식베이스 구축 단계;로 이루어진다.

<24> 또한, 상기 지식베이스 설계단계는, 각 표제어에 대하여, 백과사전의 여러 범주에 공통되는 속성에 대한 공통속성 템플릿들과, 백과사전의 개별 범주에 특징적인 속성에 대한 개별속성 템플릿들로 구성하는 것이 바람직하다.

<25> 또한, 상기 지식베이스 설계단계는, 유사의미를 갖는 속성들을 하나의 대표속성으로 통합하여 관리하고 이들의 세부적 의미는 별도의 세분류 필드에 구분하여 정의하는 것이 바람직하다.

<26> 또한, 상기 비 구조정보 추출 단계는, 학습용 코퍼스(Corpus)의 각 문장을 토큰열로 변환한 후 속성 태깅 토큰을 대상으로 의존관계를 인식하여 학습데이터를 생성하고, 최대 엔트로피 모델을 통해 이들 학습데이터를 학습하는 단계와, 백과사전 본문의 각 문장을 토큰열로 변환하고 추출대상 토큰들에 대한 의존관계를 인식한 후 그 인식결과에 상기 학습 결과 및 상기 최대 엔트로피 모델을 적용하여 각 추출대상 토큰에 대한 속성이름 및 속성값을 파악 추출하는 단계로 이루어지는 것이 바람직하다.

【발명의 구성 및 작용】

<27> 백과사전 질의응답 시스템의 지식베이스를 반자동으로 구축하는 목적을 달성하기 위하여 본 발명은 크게 다음과 같은 특징을 갖는다.

<28> 첫째, 백과사전으로부터 추출되는 정보를 저장할 지식베이스 구조를 설계함에 있어서, 지식베이스의 구조를 백과사전의 모든 범주에 공통적인 속성을 포함하는 템플릿과 각각의 개별 범주에 특징적인 속성을 포함하는 템플릿으로 구성한다.

<29> 둘째, 일반적으로 백과사전의 개요 정보가 정형화된 형식으로 기술되는 것을 고려하여, 백과사전의 개요 정보에서 속성이름과 속성값을 구조정보추출 방법을 적용하여 자동 추출한다.

<30> 셋째, 백과사전의 본문에 대해서는 학습 및 추출 단계를 통해 속성이름과 속성값을 비구조 정보추출 방법으로 추출한다. 즉, 백과사전 속성 태깅 코퍼스로부터 의존규칙을 이용하여 현재 토큰의 가장 가까운 의존소와 지배소를 좌우 문맥으로 하는 통계 모델로의 학습을 수행한다. 그리고, 이와 같이 학습된 통계 모델을 이용하여 백과사전 본문으로부터 속성값을 자동으로 추출한다. 특히, 상기 통계모델로서 최대 엔트로피 모델을 적용한다.

<31> 이하, 본 발명에 따른 실시예를 첨부한 도면을 참조하여 상세히 설명하기로 한다.

<32> 일반적으로 백과사전의 각 표제어는 개요 정보와 본문으로 구성된다. 개요 정보는 속성이름과 속성값으로 구성된 정형화된 일정한 형식을 띄고 있으며, 백과사전 본문은 비정형의 자연어 문장으로 기술된다.

<33> 도 1은 본 발명에서 주요 정보추출 대상으로 삼고 있는 백과사전에 대한 예를 보여주고 있다.

<34> 도 1에 보여지는 바와 같이, 백과사전은 표제어(100), 개요 정보(101), 및 본문(105)과 표제어가 속해 있는 범주(106) 정보로 구성된다. 또한, 개요 정보는 표제어를 설명하는 뜻풀이(102)와 속성이름(103)과 속성값(104)으로 구성되며, 본문(105)은 개요 정보와 달리 자유로운 형식의 자연어 문장으로 기술되어 있다. 또한, 범주(106)는 표제어가 속한 백과사전의 범주를 말하며 하나의 표제어가 여러 개의 범주에 속하기도 한다.

<35> 한편, 도 2는 본 발명에 따른 백과사전 지식베이스 구축 시스템에 대한 개략적인 개념도이다.

<36> 도 2에 도시된 바와 같이, 백과사전의 원문이 입력되면 구조정보 추출 모듈(201)은 백과사전의 개요 정보내의 표제어와 속성이름 및 속성값을 추출하고 이를 지식베이스(203)에 저장한다.

<37> 이와 같이 입력 백과사전 문서에서 개요 정보가 모두 추출되면, 비 구조정보 추출 모듈(202)은 백과사전의 본문에서 추출 대상이 되는 속성값을 자동으로 인식하여 지식베이스에 저장한다. 이때, 개요 정보에서 빠진 속성들도 함께 인식하여 저장한다.

<38> 또한, 지식베이스의 상세한 구조(204)는 도 2의 하단에 제시된 바와 같다. 즉, 지식베이스(203)는 표제어들(205)과 다수의 템플릿(206)으로 구성되어 있으며, 각 템플릿(206)은 여러 개의 속성(207)으로 구성된다. 그리고, 백과사전의 개요 정보(101)와 본문(105)으로부터 추출한 속성값들은 해당 템플릿의 해당 속성에 저장된다.

<39> 한편, 상기 템플릿(206)에 대한 일 실시예가 도 3에 제시되어 있다.

<40> 도 3은 백과사전 범주 체계 중 '인물'에 해당하는 템플릿에 대한 예이며, '인물' 범주에 해당하는 모든 표제어들이 공통적으로 포함하고 있는 인물 공통 속성들과 개별 범주에 특징적인 인물 개별 속성들로 구성된다.

<41> 예를 들면, '출생' 템플릿은 '출생장소', '출생일', '국적', '본관'과 같이 자주 등장하는 개념이 속성으로 정의되어 '출생' 템플릿을 구성한다. 이 외에 '명칭', '사망', '활동', '수상' 등의 주요 인물 공통 개념이 템플릿의 형태로 정의되고 각 템플릿은 여러 개의 관련 속성들로 구성됨으로써 백과사전의 지식을 개념별로 체계적으로 관리할 수 있게 된다. 이들 인물 공통 속성 외에 개별 범주별로 특징적인 개념들('구조물', '개발', '등단' 등)이 마찬가지로 템플릿의 형태로 정의되며 각각의 템플릿은 유사 개념의 다수의 속성들을 포함한다.

<42> 이와 같은 템플릿을 정의할 때 속성의 개념을 너무 세부적으로 정하게 되면, 하나의 템플릿이 많은 속성을 포함하게 되어 질의응답 시스템의 정답 검색 시간을 증가시키실 수 있으며, 일부 표제어들에게는 해당하지 않는 속성들까지 정의하게 됨으로써 공간의 낭비를 초래할 수 있다.

<43> 이러한 문제점을 해소하기 위해 본 발명은 세분화된 속성들을 대표적인 속성으로 통합하여 저장하고 이들의 세부적인 의미를 '세분류' 필드에 저장한다.

<44> 도 4는 '명칭' 템플릿에 세분류를 포함하고 있는 예이다. 각 인물 표제어의 '본명'이나 '자', '호', '별명'등을 모두 함께 '별칭'속성으로 대표 저장하고, '세분류' 필드에서 각각의 속성값이 '자', '호', '별명' 등인지를 구별한다. 이와 같이 세분화된 의미를 버리지 않고 저장함으로써 본 발명은 질의응답 시스템의 정답 검색 시간을 빠르게 하면서도 속성값의 원래 의미를 잃지 않게 되는 장점이 있다.

<45> 한편, 백과사전의 정보를 추출하는 방법은 개요 정보 추출을 위한 구조 정보추출 방법과 본문에서의 정보추출을 위한 비구조 정보추출 방법으로 나뉜다.

<46> 먼저, 도 5를 참조하여 구조정보를 추출하는 과정을 설명하도록 한다.

<47> 백과사전의 개요 정보는 일반적으로 정형화된 형식을 취하기 때문에 속성이름과 속성값의 위치가 고정되어 있다. 예를 들어, "속성이름 : 속성값"과 같은 형식으로 기술되어 있다면 콜론(':')등의 구분자를 중심으로 앞의 문자열을 속성이름으로 추출한다.(S501)

<48> 그리고, 그 추출된 속성이름이 지식베이스 템플릿의 속성 목록에 있는 유효한 속성이름인지 여부를 검사하여, 유효하지 않은 속성이름이면 다음 속성이름을 추출하는 과정을 반복하게 된다.(S502)

<49> 유효한 속성이름이면 콜론 뒤의 문자열을 속성값으로 추출한다.(S503) 이때, 추출된 속성값 중에는 하나의 속성값이 아니라 여러 개의 속성값이 등장하는 경우가 있다. 예를 들어, "한국청소년영화제 촬영상, 1983 백상예술대상 각본상, 1994 대종상영화제 신인감독상"와 같이 한 개의 속성값이 아닌 경우는 여러 구분자(,.;: 등의 심볼)를 기준으로 분리하여 각각의 속성값을 추출한다.(S505,S506)

<50> 한편, 백과사전 본문의 경우는 개요 정보와 달리 비정형의 자연어 텍스트로 기술되어 있기 때문에 자연어 분석 기법이나 기계학습 방법 등이 요구된다.

<51> 본 발명에 의한 접근방법은 언어 분석에 사용되는 의존규칙과 통계적인 기계학습 방법을 혼합함으로써 다양하고 복잡한 자연어 텍스트에 강건한(robust) 특징을 갖고 있다.

<52> 한국어는 비교적 어순이 자유롭고 격 조사 및 어미의 사용이 매우 발달되어 있는 언어이다. 또한 문맥으로 파악할 수 있으면 주어나 목적어 등과 같은 필수적인 문장 요소까지도 생략

가능하다. 이러한 한국어의 특징으로 인해 한국어의 구문 분석에는 의존 문법(dependency grammar)을 많이 이용한다. 의존 문법은 단어 사이의 의존 관계에 중심을 두는 문법이다. 즉, 문장내의 단어와 단어사이의 의존 관계 즉 수식 관계의 특성에 기반을 두고 있다.

<53> 따라서, 본 발명은 상기한 의존문법에 기반한 의존규칙을 이용하여 학습 말뭉치(Corpus)로부터 학습데이터를 추출하고 통계적인 기계학습을 통해 사실 정보를 추출한다.

<54> 또한, 본 발명에서 학습데이터를 추출하기 위한 의존규칙은 어절간의 의존관계만을 파악한다.

<55> 의존관계에 있는 두 어절에 대해 X가 Y에 의존할 때, X를 '의존소'라 하고 Y를 '지배소'라 한다. 한국어의 어절은 여러 형태소로 결합되어 있기 때문에 어절이 지배소일 때와 의존소일 때에 의존관계에 영향을 주는 형태소가 다르다. 즉, 지배소일 때에는 실질형태소(어간)가 중요한 역할을 하며 의존소일 때에는 마지막의 형식형태소(어미, 조사)가 중요한 역할을 한다.

<56> 의존관계를 파악하기 위한 규칙에 대해 간단히 살펴보면 다음과 같다.

<57> 첫째, 격 조사에 의한 의존관계로서, 의존소가 주격, 목적격, 부사격의 경우, 지배소는 가장 가까운 용언에 해당하고 의존소가 관형격, 접속격인 경우, 지배소는 가장 가까운 명사에 해당한다. 도 7에는 각 격 조사에 대한 격 정보가 보여지고 있다.

<58> 둘째, 조사가 생략된 인접한 명사나 개체명에 대해서는 선행하는 명사 또는 개체명이 의존소이고, 후행하는 명사 또는 개체명이 지배소가 된다.

<59> 셋째, 백과사전 본문의 경우 다른 문서와는 다르게 기호가 많이 등장하는 특징이 있다. 예를 들어, "저서로 《Studies in Ethnomethodology》(1967)가 있다."와 같이 저서명 주위의 '

. 《', '》'나 년도 주위의 '(', ')'등의 쓰임새가 많다. 따라서 격 조사가 붙어 있지 않는 명사들 중에서 다음 토큰이 기호인 경우 지배소는 용언으로 한다.

<60> 상기된 바와 같은 의존규칙들을 바탕으로 한국어의 지배소 후위의 원칙을 적용하게 되면, 현재 토큰을 중심으로 가장 인접한 좌측의 의존소와 우측의 지배소를 찾아서 이들을 좌우 문맥으로 하는 학습데이터를 추출할 수 있다.

<61> 이와 같이 학습 자질(Feature)을 추출하는 것은, 기존의 영어권 언어에서와 같이 현재 토큰의 주변 문맥을 학습 자질로 추출하는 방법이 어순이 비교적 자유로운 한국어의 특성에 맞지 않기 때문이다. 또한, 어느 언어에나 단어들 사이의 의존관계는 존재하기 때문에 이러한 학습 자질 추출 방법은 한국어뿐만 아니라 영어권 언어에도 적용될 수 있다.

<62> 또한, 본 발명은 상기 방법으로 추출된 학습데이터를 최대 엔트로피 모델(Maximum Entropy Model)을 이용하여 학습한다. 최대 엔트로피 모델은 어떤 종류의 자질이든지 모두 수용할 수 있으며, 자질을 학습하는 방법보다는 어떠한 자질을 사용할 지만 결정해주면 내부 알고리즘에 의해 자동으로 파라미터(parameter)가 결정되기 때문에 다양한 문제들에 동일한 학습 엔진을 재 사용할 수 있는 장점이 있다. 또한, 다른 통계 모델과는 달리 자질이 많이 추가되더라도 모델의 안정성이 보장된다.

<63>

$$E[f_j] = \tilde{E}[f_j], \qquad 1 \le j \le k$$

$$E[f_j] = \sum_{x,y} p(x,y) f_j(x,y)$$

$$\tilde{E}[f_j] = \sum_{i=1}^{n} \tilde{p}(x_i,y_i) f_j(x_i,y_j)$$

<64> 상기 식은 최대 엔트로피 모델의 제약 조건식(Constraint Equation)이다. 여기에서 f는 자질 함수를 의미하는데 본 발명에서는 의존규칙을 의미하며, 말뭉치로부터 관측되는 모든 문맥들 중에서 f를 만족하는 문맥 데이터들은 $\tilde{E}[f_j]$의 확률 분포들로 나타낸다. x는 의존규칙으로 추출될 수 있는 모든 문맥의 집합을 나타내고, y는 모든 속성 유형을 나타낸다. 그리고, n은 말뭉치의 학습 데이터에서 발견된 x와 y의 곱집합으로 얻을 수 있는 총 가지 수를 의미한다. 상기의 제약 조건식을 만족하는 확률 분포들 중에서 엔트로피가 최대가 되는 확률 분포를 찾는 것이 최대 엔트로피 원리에 의한 통계 모델이다.

<65>

$$P = \{p \mid E[f_j] = \tilde{E}[f_j], j = \{1...k\}\}$$

$$p^* = \arg\max_{p \in P} H(p)$$

$$H(p) = -\sum_{x,y} p(x,y)\log p(x,y) = -\sum_{x,y} \tilde{p}(x)p(y \mid x)\log p(y \mid x)$$

$$p(y \mid x) = \frac{1}{Z(x)}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right]$$

$$Z(x) = \sum_{y}\exp\left[\sum_{i=1}^{k}\lambda_i f_i(x,y)\right]$$

<66> 상기 수식에서 P*는 P를 만족하는 확률 분포들 중에서 엔트로피가 최대가 되는 확률 분포를 의미하고, H(p)는 엔트로피의 계산식이다. H(p)식의 p(y|x)내의 λ는 여러 자질 함수들에 대한 가중치를 의미하는데, 학습 단계에서 내부 알고리즘에 의해 자동으로 결정되는 파라미터이다. 이 파라미터를 결정하는 작업이 엔트로피를 최대로 하는 확률 분포를 찾기 위한 주요 학습 내용이다. 마지막의 Z(x)는 계산된 수치값을 정규화(normalization)하기 위한 수식이다.

<67> 이하에서는, 도 6내지 도 9를 참조하여 본 발명에 따른 백과사전의 본문으로부터 비 구조정보를 추출하는 과정에 대해 설명한다.

<68> 도 6은 본 발명에 따른 비 구조정보 추출 방법의 학습 단계와 추출 단계의 각 과정을 보여준다. 문서로부터 정보를 추출하려면 먼저 학습 단계를 거친 후에 추출 단계를 수행해야 한다.

<69> 학습 단계에서는, 개체명과 속성이 태깅된 백과사전 학습용 텍스트 코퍼스를 입력받아 형태소 분석을 수행한다.(S601,S602) 그리고, 각 문장별로 형태소 분석 결과를 토대로 어절 단위의 토큰열을 인식한다.(S603) 또한, 문장내에서 속성값으로 태깅된 토큰들을 대상으로 상기 설명된 바와 같은 의존규칙을 이용하여 지배소와 의존소의 의존관계를 인식한다.(S604) 그리고, 이들을 좌우 문맥으로 하는 학습데이터를 생성한 후 최대 엔트로피 모델로 학습한 결과와 함께 통계모델 DB(600)에 저장한다.(S605)

<70> 상기 통계 모델로는 최대 엔트로피 모델이 바람직하지만, 이에 한정되지 않고 당업자로서는 최대 엔트로피 부스팅 모델, Back-off 모델, 결정 트리 모델 등 실시 가능한 다른 모델을 적용할 수 있다.

<71> 도 8a에는 이와 같은 학습과정에 대한 예(800)를 보여주고 있다.

<72> 예를 들면, "<RG><기센대학:ORGANIZATION></RG>의 <RP>위생학 <교수:POSITION></RP>를 역임하였다."라는 개체명 및 속성이 태깅된 말뭉치의 문장을 가지고 설명하면 다음과 같다.

<73> 주어진 속성 태깅 문장의 '<RG>'는 경력을 나타내는 단체명의 속성값이 시작하는 위치를 뜻하며 '</RG>'는 속성값이 끝나는 위치를 나타낸다. 그리고 '<RP>'는 경력의 지위를 나타낸다. 또한, '<기센대학:ORGANIZATION>'은 '기센대학'이 ORGANIZATION의 개체명임을 나타내고, '<교수:POSITION>'은 '교수'라는 단어가 POSITION의 개체명임을 나타낸다.

<74> 상기한 문장은 어절 단위의 토큰열로 인식되어 (801), (802), (803)에서 볼 수 있는 바와 같이 속성값에 해당하는 현재 토큰(802)을 중심으로 좌측 의존소(801)와 우측 지배소(803)의 학습데이터를 생성한다.

<75> 또한, Tag(804)는 현재 토큰의 속성 태그를 나타낸다. 각각의 속성 태그의 의미는 도 9와 같다. 도 8의 Tag(804)는 도 9의 속성 태그에 접미어로 '_SE'(단일), '_ST'(시작), '_ED'(종료)가 붙은 값이다. 즉, 각각의 현재 토큰이 해당 속성값의 어디에 위치하고 있는 지를 나타낸 것이다. 이렇게 하는 것은 하나의 속성값이 여러 토큰으로 구성되어 있는 경우를 처리하기 위함이다.

<76> 상기 설명된 바와 같이 비 구조정보 추출을 위한 학습 단계가 완료되면, 백과사전의 본문으로부터 정보를 추출하는 추출 단계를 수행한다.

<77> 도 6을 참조하면, 정보 추출단계는 먼저 백과사전 본문에 대해 문장 단위로 형태소 분석 및 개체명 인식 과정을 수행한다.(S611)

<78> 또한, 이러한 형태소 분석 및 개체명 인식 결과를 토대로 각각의 문장을 어절 단위의 토큰열로 변환한다.(S612) 그리고, 개체명에 해당하거나 실질형태소가 명사에 해당하는 토큰을 토큰열로부터 검색하여 현재 토큰으로 지정한다. 현재 토큰을 이와 같이 제한한 이유는 대부분의 속성 태깅 대상에 해당하는 토큰이 개체명과 일반 명사들이기 때문이다.

<79> 그리고, 각 지정된 현재 토큰에 의존규칙을 적용하여 의존소와 지배소에 해당하는 좌우 문맥 토큰을 찾아낸다.(S613)

<80> 이들 토큰들은 학습단계에서 학습된 최대 엔트로피 모델의 입력이 되며 현재 토큰에 대한 속성유형들을 분류하고,(S614) 이들중에서 가장 높을 확률을 갖는 것을 현재 토큰에 대한 속성이름 및 속성값으로 추출한다.(S615)

<81> 도 8b는 "기센대학의 위생학 교수를 역임하였다."라는 예시 문장에서의 정보 추출예를 보여준다.

<82> 상기의 문장은 형태소 분석 및 개체명 인식 과정을 거쳐서 "<기센대학:ORGANIZATION>의 위생학 <교수:POSITION>을 역임하였다."와 같은 형태로 태깅된 후 어절단위의 토큰열로 변환된다. 그리고, 개체명인 '<기센대학:ORGANIZATION>', '<교수:POSITION>'과 명사인 '위생학'이 현재 토큰으로 지정되고 각각의 의존소와 지배소를 토큰열로부터 찾아서 최대 엔트로피 모델에 의해 현재 토큰의 속성 유형 분류를 하게 된다. 지시창 805는 가장 엔트로피가 큰 순서대로 각 현재 토큰에 대한 속성 유형들을 나열한 결과이다. 이들 후보 속성유형 중에서 가장 왼쪽에 위치한 태그를 최종 결과로 한다. 즉, '<기센대학:ORGANIZATION>의'는 'RG_SE'으로, '위생학'은 'RP_ST'로, '<교수:POSITION>를'은 'RP_ED'로 속성값 추정을 하고, '(경력.단체, 기센대학)', '(경력.지위, 위생학 교수)'의 속성값으로 추출하게 된다.

<83> 한편, 상기와 같이 비 구조 정보추출 과정을 거치게 되면 일반적으로 다수의 '(속성이름, 속성값)'을 얻게 되는데, 이들 중에는 상기 구조정보 추출 단계에서 이미 개요 정보로부터 추출된 속성값이 있을 수 있다.

<84> 따라서, 상기 비 구조정보로부터 추출된 속성값을 저장하기 전에 속성이름을 사용하여 지식베이스 내에 이미 해당 속성값이 존재하는지 검사한 후, 속성이름이 이미 존재한다면 비구조 정보추출로 추출된 속성값은 버려지게 되며, 존재하지 않는다면 개요 정보에는 없는 속성값이므로 추가로 보완하여 채움으로써 지식베이스를 구축하게 된다.

【발명의 효과】

<85> 상술한 바와 같이 본 발명에 따른 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법은, 백과사전 지식의 체계적인 템플릿을 기반으로 반자동으로 지식베이스를 구축함으로써 보다 가치 있고 유연한 지식 형태를 구축할 수 있으며 백과사전 질의응답시스템의 정답 제공을 빠르고 용이하게 하고, 지식베이스 구축에 따른 비용과 노력을 절감할 수 있다.

<86> 또한 본 발명에 따르면, 백과사전 개요 정보에 빠져있는 유용한 정보를 백과사전 본문을 통해 자동으로 인식, 추출하여 개요 정보의 형태로 보완하여 구축해줌으로써 백과사전의 지식베이스 완성도를 높여준다.

<87> 또한 본 발명에 따르면, 기존의 비구조 정보추출을 위한 접근방법에서 단순히 주변 문맥을 학습하는 것과는 달리 문장내의 의존관계를 분석하여 학습함으로써 비교적 어순이 자유로운 한국어에 매우 적합하여 성능을 향상시켜 줄뿐만 아니라 영어권 언어에서도 쉽게 적용 가능하다.

<88> 또한 본 발명에 따르면, 다양하고 많은 학습 자질을 학습해야 하는 비구조 정보추출을 위해 최대 엔트로피 모델을 적용함으로써 대량의 학습 데이터를 학습하더라도 안정된 성능을 보장할 수 있다.

<89> 이상에서 설명한 것은 본 발명에 따른 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법을 실시하기 위한 하나의 실시예에 불과한 것으로서, 본 발명은 상기한 실시예에 한정되지 않고, 이하의 특허청구의 범위에서 청구하는 본 발명의 요지를 벗어남이 없이 당해 발

명이 속하는 분야에서 통상의 지식을 가진 자라면 누구든지 다양한 변경 실시가 가능한 범위까지 본 발명의 기술적 정신이 있다고 할 것이다.

【특허청구범위】

【청구항 1】

각 표제어에 대해 다수의 템플릿들과 각 템플릿에 대한 다수의 관련 속성들로 지식베이스의 구조를 설계하는 지식베이스 설계단계;

백과사전의 개요정보로부터 표제어와 그 속성이름 및 속성값들을 추출하는 구조정보 추출 단계;

백과사전의 본문으로부터 그 표제어에 대한 속성이름 및 속성값들을 추출하는 비 구조정보 추출 단계; 및

각 표제어별로 상기 추출된 구조정보 및 비 구조정보를 지식베이스의 해당 템플릿 및 해당 속성에 저장하는 지식베이스 구축 단계;로 이루어지는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 2】

제 1항에 있어서, 상기 지식베이스 설계단계는, 각 표제어에 대하여, 백과사전의 여러 범주에 공통되는 속성에 대한 공통속성 템플릿들과, 백과사전의 개별 범주에 특징적인 속성에 대한 개별속성 템플릿들로 구성하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 3】

제 1항 또는 제 2항에 있어서, 상기 지식베이스 설계단계는,

유사의미를 갖는 속성들을 하나의 대표속성으로 통합하여 관리하고 이들의 세부적 의미는 별도의 세분류 필드에 구분하여 정의하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 4】

제 1항에 있어서, 상기 구조정보 추출 단계는,

개요정보의 정형화된 형식을 인지하여 속성이름과 속성값을 추출하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 5】

제 4항에 있어서, 상기 구조정보 추출 단계는,

개요정보의 정형화된 형식을 통해 먼저 속성이름을 추출하여 그 속성이름이 지식베이스 템플릿의 속성목록에 있는 유효한 속성인지를 검사한 후, 유효 속성일 경우에만 그 속성값을 추출하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 6】

제 4항 또는 제 5항에 있어서, 상기 구조정보 추출 단계는,

상기 추출된 속성이름에 대해 여러 개의 속성값이 있을 경우, 그 표시된 구분자를 통해 각각을 분리하여 추출하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 7】

제 1항에 있어서, 상기 비 구조정보 추출 단계는,

학습용 코퍼스(Corpus)의 각 문장을 토큰열로 변환한 후 속성 태깅 토큰을 대상으로 의존관계를 인식하여 학습데이터를 생성하고, 소정의 통계모델을 통해 이들 학습데이터를 학습하는 단계와,

백과사전 본문의 각 문장을 토큰열로 변환하고 추출대상 토큰들에 대한 의존관계를 인식한 후 그 인식결과에 상기 학습 결과 및 상기 통계모델을 적용하여 각 추출대상 토큰에 대한 속성이름 및 속성값을 파악 추출하는 단계로 이루어지는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 8】

제 7항에 있어서, 상기 학습 단계는,

개체명과 속성이 태깅된 백과사전 학습용 코퍼스(Corpus)에 대해 형태소 분석을 수행하고, 각 문장별로 어절단위 토큰열을 인식하는 단계와,

토큰열에서 속성값이 태깅된 토큰을 대상으로 소정의 의존규칙을 적용하여 대상 토큰에 대한 지배소와 의존소의 의존관계를 인식하는 단계와,

각 대상 토큰의 지배소와 의존소를 좌우 문맥으로 하여 학습데이터를 생성한 후, 소정의 통계모델로 학습한 결과를 저장하는 단계로 이루어지는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 9】

제 7항에 있어서, 상기 추출 단계는,

백과사전 본문에 형태소 분석 및 개체명 인식을 수행하여 각 문장을 어절 단위의 토큰열로 변환하는 단계와,

상기 토큰열에서 개체명이거나 실질 형태소가 명사인 토큰을 추출대상 토큰으로 지정하는 단계와,

각 지정 추출대상 토큰에 소정의 의존규칙을 적용하여 지배소와 의존소의 좌우 문맥 토큰을 인식하는 단계와,

상기 추출대상 토큰 및 그 좌우 문맥 토큰을 상기 학습결과 및 상기 통계모델에 적용하여 추출대상 토큰에 대한 속성유형들을 분류하고, 이중 확률이 가장 높은 것을 추출대상 토큰의 속성이름 및 속성값으로 추출하는 단계로 이루어지는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 10】

제 8항 또는 제 9항에 있어서, 상기 의존관계 인식을 위해 적용되는 의존 규칙은,

격 조사에 의한 의존관계로서, 의존소가 주격, 목적격, 부사격일 경우 그 지배소는 가장 가까운 용언으로 하고, 의존소가 관형격, 접속격인 경우 그 지배소는 가장 가까운 명사가 되는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 11】

제 8항 또는 제 9항에 있어서, 상기 의존관계 인식을 위해 적용되는 의존 규칙은,

조사가 생략된 인접한 명사나 개체명에 대해서는 선행하는 명사 또는 개체명이 의존소이고, 후행하는 명사 또는 개체명이 지배소가 되는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【청구항 12】

제 8항 또는 제 9항에 있어서, 상기 의존관계 인식을 위해 적용되는 의존 규칙은,

격 조사가 붙어 있지 않는 명사들에 대해 그 주위 토큰이 기호일 경우 문장의 용언을 그 지배소로 하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법

.

【청구항 13】

제 7항내지 제 9항중 어느 한항에 있어서, 상기 통계모델로서, 최대 엔트로피 모델을 사용하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

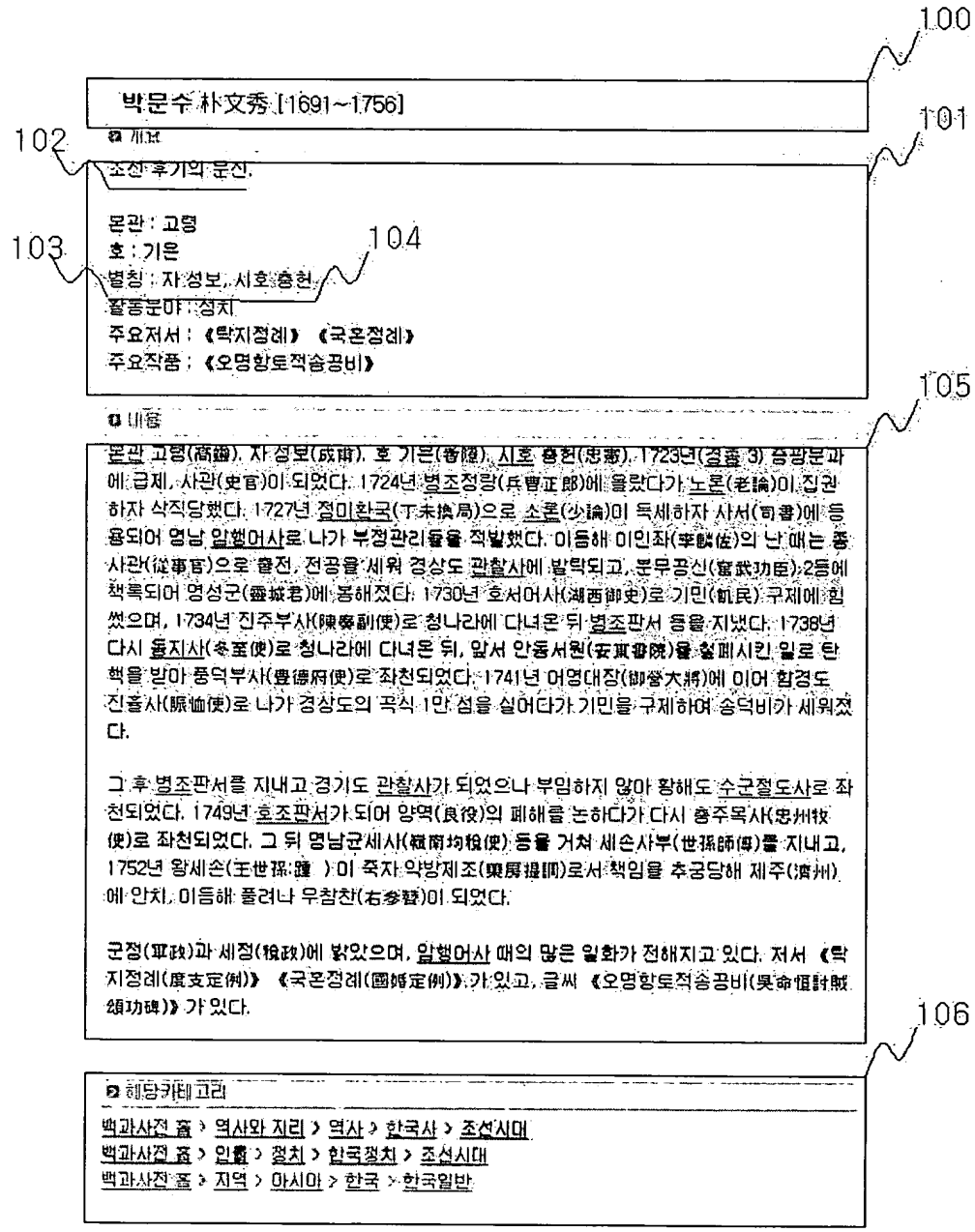
【청구항 14】

제 1항 있어서, 상기 지식베이스 구축 단계는,

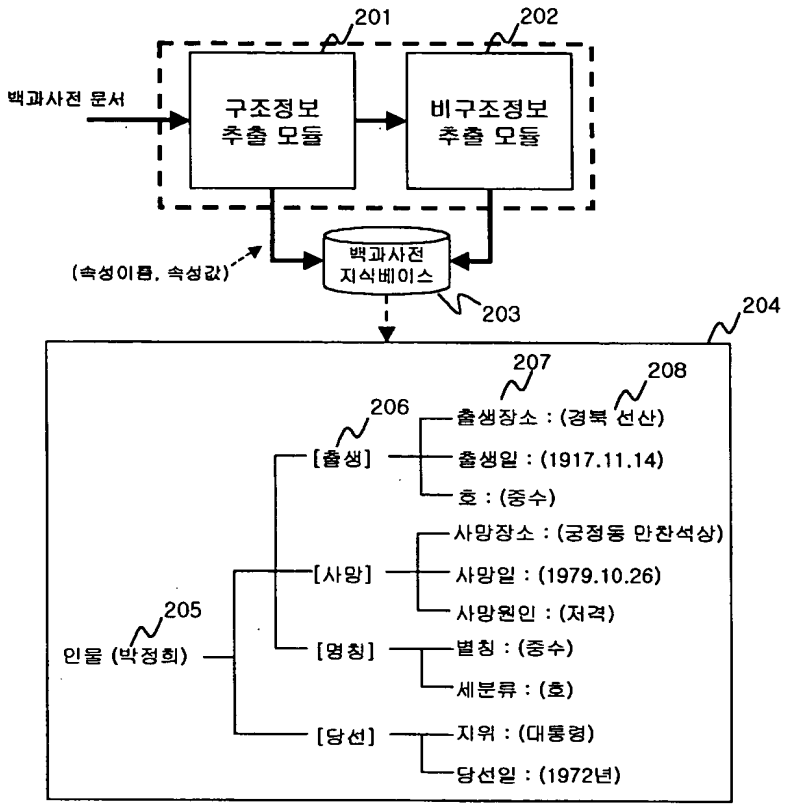
상기 구조정보 추출단계를 통해 추출된 속성이름 및 속성값으로 지식베이스를 먼저 구축한 후, 상기 비 구조정보로 추출된 속성이름 및 속성값은 지식베이스내에 그 표제어에 대해 동일 속성값의 존재여부에 따라 보완적으로 저장하는 것을 특징으로 하는 백과사전 질의응답 시스템의 지식베이스 반자동 구축 방법.

【도면】

【도 1】


100
박문수 朴文秀 [1691~1756]
101
102
조선 후기의 문신.
본관 : 고령
호 : 기은
별칭 : 자성보, 시호 충헌
활동분야 : 정치
주요저서 : 《탁지정례》 《국혼정례》
주요작품 : 《오명항토적송공비》
103
104
105
내용
본관 고령(高靈). 자 성보(成甫). 호 기은(耆隱). 시호 충헌(忠憲). 1723년(경종 3) 증광문과에 급제, 사관(史官)이 되었다. 1724년 병조정랑(兵曹正郎)에 올랐다가 노론(老論)이 집권하자 삭직당했다. 1727년 정미환국(丁未換局)으로 소론(少論)이 득세하자 사서(司書)에 등용되어 영남 암행어사로 나가 부정관리들을 적발했다. 이듬해 이인좌(李麟佐)의 난 때는 종사관(從事官)으로 출전, 전공을 세워 경상도 관찰사에 발탁되고, 분무공신(奮武功臣) 2등에 책록되어 영성군(靈城君)에 봉해졌다. 1730년 호서어사(湖西御史)로 기민(飢民) 구제에 힘썼으며, 1734년 진주부사(陳奏副使)로 청나라에 다녀온 뒤 병조판서 등을 지냈다. 1738년 다시 동지사(冬至使)로 청나라에 다녀온 뒤, 앞서 안동서원(安東書院)을 철폐시킨 일로 탄핵을 받아 풍덕부사(豊德府使)로 좌천되었다. 1741년 어영대장(御營大將)에 이어 함경도 진휼사(賑恤使)로 나가 경상도의 곡식 1만 섬을 실어다가 기민을 구제하여 송덕비가 세워졌다.
그 후 병조판서를 지내고 경기도 관찰사가 되었으나 부임하지 않아 황해도 수군절도사로 좌천되었다. 1749년 호조판서가 되어 양역(良役)의 폐해를 논하다가 다시 충주목사(忠州牧使)로 좌천되었다. 그 뒤 영남균세사(嶺南均稅使) 등을 거쳐 세손사부(世孫師傅)를 지내고, 1752년 왕세손(王世孫:正)이 죽자 악방제조(樂房提調)로서 책임을 추궁당해 제주(濟州)에 안치, 이듬해 풀려나 우참찬(右參贊)이 되었다.
군정(軍政)과 세정(稅政)에 밝았으며, 암행어사 때의 많은 일화가 전해지고 있다. 저서 《탁지정례(度支定例)》 《국혼정례(國婚定例)》가 있고, 글씨 《오명항토적송공비(吳命恒討賊頌功碑)》가 있다.
106
해당카테고리
백과사전 홈 > 역사와 지리 > 역사 > 한국사 > 조선시대
백과사전 홈 > 인물 > 정치 > 한국정치 > 조선시대
백과사전 홈 > 지역 > 아시아 > 한국 > 한국일반

【도 2】



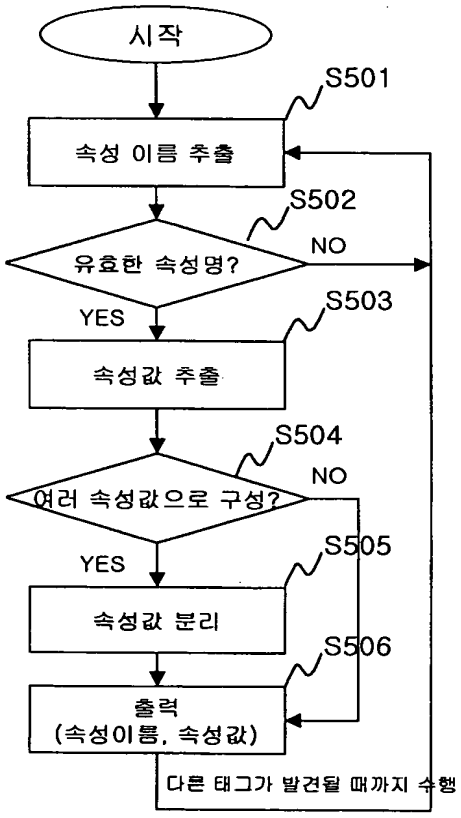
【도 3】

	템플릿명	속성명
인물 공통 속성	출생	출생장소
		출생일
		국적
		본관
	명칭	별칭
	사망	사망일
	활동	활동분야
	수상	수상명
		수상일
인물 개별 속성	구조물	건축물
		건축년도
	개발	개발품
		개발일
	등단	등단작
		등단공간
		등단일

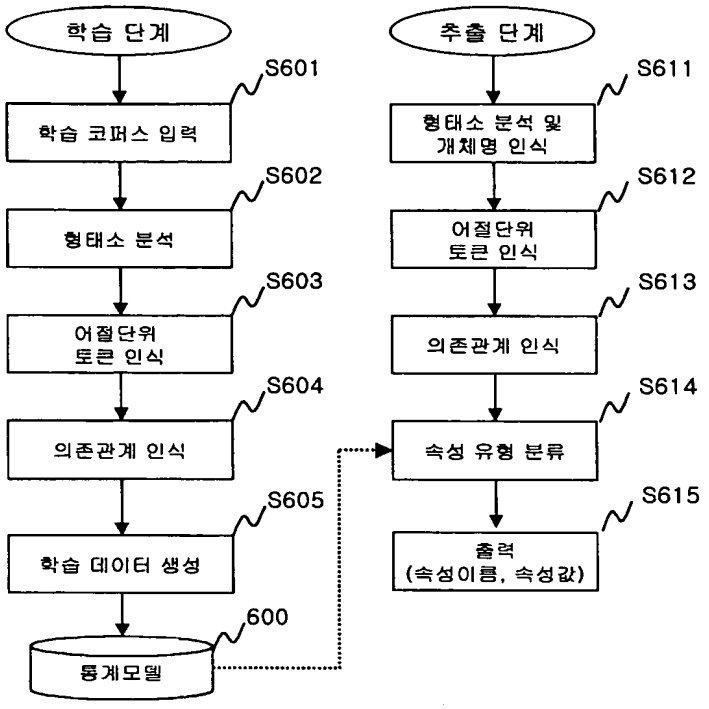
【도 4】

템플릿 이름	속성 이름	속성값의 예제
명칭	별칭	'영서' '군섭' '월봉' '정강' '게이터'
	세분류	자/호/별명/본명/ 별칭

【도 5】



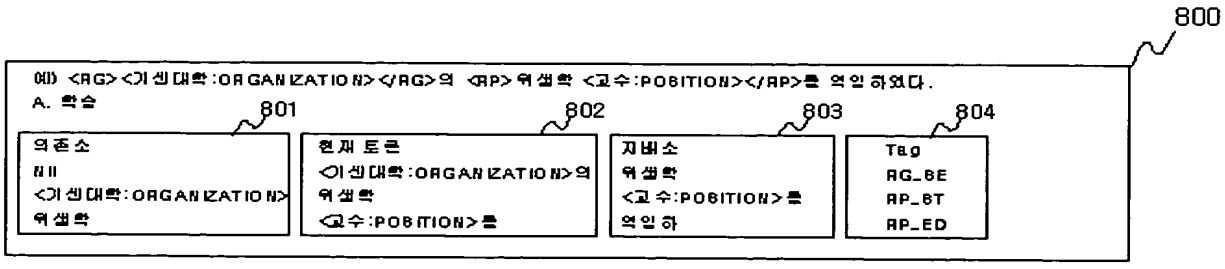
【도 6】



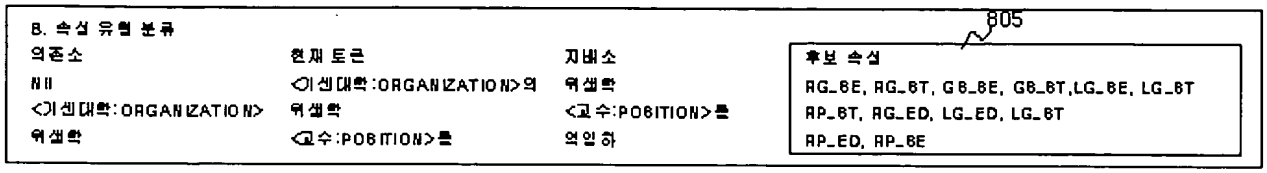
【도 7】

조사	주격	목적격	부사격	관형격	접속격
은/는	O	O	O	X	X
이/가	O	X	X	X	X
도	O	O	O	X	X
을/를	X	O	X	X	X
로/으로	X	X	O	X	X
서/에서	X	X	O	X	X
와/과	O	O	O	O	O
의	X	X	X	O	

【도 8a】



【도 8b】



【도 9】

템플릿 이름	속성 이름	속성태그
학력	졸업학교	GS
경력	단체	RG
경력	직위	RP
당선	정당명	LG
당선	지위	LP